IN THE SPECIFICATION

Please insert the following paragraphs after page 3, line 23:

FIG. 14 shows an example of set of concepts that can form a directed set.

FIG. 15 shows a directed set constructed from the set of concepts of FIG. 14 in a preferred embodiment of the invention.

FIGs. 16A-16G show eight different chains in the directed set of FIG. 15 that form a basis for the directed set.

FIG. 17 shows data structures for storing a directed set, chains, and basis chains, such as the directed set of FIG. 14, the chains of FIG. 15, and the basis chains of FIGs. 16A-16G.

Please amend the paragraph beginning on page 3, line 27, as follows:

A semantic abstract representing the content of the document can be constructed as a set of vectors within the topological vector space. ~~(~~The construction of state vectors in a topological vector space is described in U.S. Patent application Serial No. 09/512,963, titled "CONSTRUCTION, MANIPULATION, AND COMPARISON OF A MULTI-DIMENSIONAL SEMANTIC SPACE," filed February 25, 2000, incorporated by reference herein and referred to as "the Construction application."~~)~~ The following text is copied from that application:

> At this point, a concrete example of a (very restricted) lexicon is in order. FIG. 3 shows a set of concepts, including "thing" 1405, "man" 1410, "girl" 1412, "adult human" 1415, "kinetic energy" 1420, and "local action" 1425. "Thing" 1405 is the maximal element of the set, as every other concept is a type of "thing." Some concepts, such as "man" 1410 and "girl" 1412 are "leaf concepts," in the sense that no other concept in the set is a type of "man" or "girl." Other concepts, such as "adult human" 1415, "kinetic energy" 1420, and "local action" 1425 are "internal concepts," in the sense that they are types of other concepts (e.g., "local action" 1425 is a type of "kinetic energy" 1420) but there are other concepts that are types of these concepts (e.g., "man" 1410 is a type of "adult human" 1415).
>
> FIG. 4 shows a directed set constructed from the concepts of FIG. 3. For each concept in the directed set, there is at least one chain extending from maximal element "thing" 1405 to the concept. These chains are composed of directed links, such as links 1505, 1510, and 1515, between pairs of concepts. In the directed set of FIG. 4, every chain from maximal element "thing" must pass through either "energy" 1520 or "category" 1525. Further, there can be more than one chain extending from maximal

element "thing" 1405 to any concept. For example, there are four chains extending from "thing" 1405 to "adult human" 1415: two go along link 1510 extending out of "being" 1535, and two go along link 1515 extending out of "adult" 1545.

Some observations about the nature of FIG. 4:

- First, the model is a *topological space.*
- Second, note that *the model is not a tree.* In fact, it is an example of a *directed set.* For example, concepts "being" 1530 and "adult human" 1415 are types of multiple concepts higher in the hierarchy. "Being" 1530 is a type of "matter" 1535 and a type of "behavior" 1540; "adult human" 1415 is a type of "adult" 1545 and a type of "human" 1550.
- Third, observe that the relationships expressed by the links are indeed relations of hyponymy.
- Fourth, note particularly – but without any loss of generality – that "man" 1410 maps to both "energy" 1520 and "category" 1525 (via composite mappings) which in turn both map to "thing" 1405; i.e., the (composite) relations are multiple valued and induce a partial ordering. These multiple mappings are natural to the meaning of things and critical to semantic characterization.
- Finally, note that "thing" 1405 is *maximal*; indeed, "thing" 1405 is the *greatest* element of *any* quantization of the lexical semantic field (subject to the premises of the model).

Metrizing **S**

FIGs. 5A-5G show eight different chains in the directed set that form a basis for the directed set. FIG. 5A shows chain 1605, which extends to concept "man" 1410 through concept "energy" 1520. FIG. 5B shows chain 1610 extending to concept "iguana." FIG. 5C shows another chain 1615 extending to concept "man" 1410 via a different path. FIGs. 5D-5G show other chains.

FIG. 13 shows a data structure for storing the directed set of FIG. 3, the chains of FIG. 4, and the basis chains of FIGs. 5A-5G. In FIG. 13, concepts array 1705 is used to store the concepts in the directed set. Concepts array 1705 stores pairs of elements. One element identifies concepts by name; the other element stores numerical identifiers 1706. For example, concept name 1707 stores the concept

"dust," which is paired with numerical identifier "2" 1708. Concepts array 1705 shows 9 pairs of elements, but there is no theoretical limit to the number of concepts in concepts array 1705. In concepts array 1705, there should be no duplicated numerical identifiers 1706. In FIG. 13, concepts array 1705 is shown sorted by numerical identifier 1706, although this is not required. When concepts array 1705 is sorted by numerical identifier 1706, numerical identifier 1706 can be called the *index* of the concept name.

Maximal element (ME) 1710 stores the index to the maximal element in the directed set. In FIG. 13, the concept index to maximal element 1710 is "6," which corresponds to concept "thing," the maximal element of the directed set of FIG. 4.

Chains array 1715 is used to store the chains of the directed set. Chains array 1715 stores pairs of elements. One element identifies the concepts in a chain by index; the other element stores a numerical identifier. For example, chain 1717 stores a chain of concept indices "6", "5", "9", "7", and "2," and is indexed by chain index "1" (1718). (Concept index 0, which does not occur in concepts array 1705, can be used in chains array 1715 to indicate the end of the chain. Additionally, although chain 1717 includes five concepts, the number of concepts in each chain can vary.) Using the indices of concepts array 1705, this chain corresponds to concepts "thing," "energy," "potential energy," "matter," and "dust." Chains array 1715 shows one complete chain and part of a second chain, but there is no theoretical limit to the number of chains stored in chain array 1715. Observe that, because maximal element 1710 stores the concept index "6," every chain in chains array 1715 should begin with concept index "6." Ordering the concepts within a chain is ultimately helpful in measuring distances between the concepts. However concept order is not required. Further, there is no required order to the chains as they are stored in chains array 1715.

Basis chains array 1720 is used to store the chains of chains array 1715 that form a basis of the directed set. Basis chains array 1720 stores chain indices into chains array 1715. Basis chains array 1720 shows four chains in the basis (chains 1, 4, 8, and 5), but there is no theoretical limit to the number of chains in the basis for the directed set.

Euclidean distance matrix 1725A stores the distances between pairs of concepts in the directed set of FIG. 4. (How distance is measured between pairs of concepts in the directed set is discussed below. But in short, the concepts in the

directed set are mapped to state vectors in multi-dimensional space, where a state vector is a directed line segment starting at the origin of the multi-dimensional space and extending to a point in the multi-dimensional space.) The distance between the end points of pairs of state vectors representing concepts is measured. The smaller the distance is between the state vectors representing the concepts, the more closely related the concepts are. Euclidean distance matrix 1725A uses the indices 1706 of the concepts array for the row and column indices of the matrix. For a given pair of row and column indices into Euclidean distance matrix 1725A, the entry at the intersection of that row and column in Euclidean distance matrix 1725A shows the distance between the concepts with the row and column concept indices, respectively. So, for example, the distance between concepts "man" and "dust" can be found at the intersection of row 1 and column 2 of Euclidean distance matrix 1725A as approximately 1.96 units. The distance between concepts "man" and "iguana" is approximately 1.67, which suggests that "man" is closer to "iguana" than "man" is to "dust." Observe that Euclidean distance matrix 1725A is symmetrical: that is, for an entry in Euclidean distance matrix 1725A with given row and column indices, the row and column indices can be swapped, and Euclidean distance matrix 1725A will yield the same value. In words, this means that the distance between two concepts is not dependent on concept order: the distance from concept "man" to concept "dust" is the same as the distance from concept "dust" to concept "man."

Angle subtended matrix 1725B is an alternative way to store the distance between pairs of concepts. Instead of measuring the distance between the state vectors representing the concepts (see below), the angle between the state vectors representing the concepts is measured. This angle will vary between 0 and 90 degrees. The narrower the angle is between the state vectors representing the concepts, the more closely related the concepts are. As with Euclidean distance matrix 1725A, angle subtended matrix 1725B uses the indices 1706 of the concepts array for the row and column indices of the matrix. For a given pair of row and column indices into angle subtended matrix 1725B, the entry at the intersection of that row and column in angle subtended matrix 1725B shows the angle subtended the state vectors for the concepts with the row and column concept indices, respectively. For example, the angle between concepts "man" and "dust" is approximately 51 degrees, whereas the angle between concepts "man" and "iguana" is approximately 42 degrees.

This suggests that "man" is closer to "iguana" than "man" is to "dust." As with Euclidean distance matrix 1725A, angle subtended matrix 1725B is symmetrical.

Not shown in FIG. 13 is a data structure component for storing state vectors (discussed below). As state vectors are used in calculating the distances between pairs of concepts, if the directed set is static (i.e., concepts are not being added or removed and basis chains remain unchanged), the state vectors are not required after distances are calculated. Retaining the state vectors is useful, however, when the directed set is dynamic. A person skilled in the art will recognize how to add state vectors to the data structure of FIG. 13.

Although the data structure for concepts array 1705, maximal element 1710 chains array 1715, and basis chains array 1720 in FIG. 13 are shown as arrays, a person skilled in the art will recognize that other data structures are possible. For example, concepts array could store the concepts in a linked list, maximal element 1710 could use a pointer to point to the maximal element in concepts array 1705, chains array 1715 could use pointers to point to the elements in concepts array, and basis chains array 1720 could use pointers to point to chains in chains array 1715. Also, a person skilled in the art will recognize that the data in Euclidean distance matrix 1725A and angle subtended matrix 1725B can be stored using other data structures. For example, a symmetric matrix can be represented using only one half the space of a full matrix if only the entries below the main diagonal are preserved and the row index is always larger than the column index. Further space can be saved by computing the values of Euclidean distance matrix 1725A and angle subtended matrix 1725B "on the fly" as distances and angles are needed.

Returning to FIGs. 5A-5G, how are distances and angles subtended measured? The chains shown in FIGs. 5A-5G suggest that the relation between any node of the model and the maximal element "thing" 1405 can be expressed as any one of a set of *composite* functions; one function for each chain from the minimal node μ to "thing" 1405 (the n^{th} predecessor of μ along the chain):

$$f: \mu \Rightarrow thing = f_1 \circ f_2 \circ f_3 \circ \cdots \circ f_n$$

where the chain connects $n + 1$ concepts, and f_j: links the $(n - j)^{th}$ predecessor of μ with the $(n + 1 - j)^{th}$ predecessor of μ, $1 \leq j \leq n$. For example, with reference to FIG. 5A, chain 1605 connects nine concepts. For chain 1605, f_1 is link 1605A, f_2 is link 1605B, and so on through f_8 being link 1605H.

Consider the set of all such functions for all minimal nodes. Choose a countable subset $\{f_k\}$ of functions from the set. For each f_k construct a function g_k: S $\Rightarrow \mathbf{I}$ [1] as follows. For $s \in$ S, s is in relation (under hyponymy) to "thing" 1405. Therefore, s is in relation to at least one predecessor of μ, the minimal element of the (unique) chain associated with f_k. Then there is a predecessor of smallest index (of μ), say the m^{th}, that is in relation to s. Define:

$$g_k(s) = (n - m) / n \qquad \text{Equation (1)}$$

This formula gives a measure of concreteness of a concept to a given chain associated with function f_k.

As an example of the definition of g_k, consider chain 1605 of FIG. 5A, for which n is 8. Consider the concept "cat" 1655. The smallest predecessor of "man" 1410 that is in relation to "cat" 1655 is "being" 1530. Since "being" 1530 is the fourth predecessor of "man" 1410, m is 4, and g_k("cat" 1655) = (8 - 4) / 8 = ½. "Iguana" 1660 and "plant" 1660 similarly have g_k values of ½. But the only predecessor of "man" 1410 that is in relation to "adult" 1545 is "thing" 1405 (which is the eighth predecessor of "man" 1410), so m is 8, and g_k("adult" 1545) = 0.

Finally, define the vector valued function φ: S $\Rightarrow \mathbb{R}^k$ relative to the indexed set of scalar functions $\{g_1, g_2, g_3, \ldots, g_k\}$ (where scalar functions $\{g_1, g_2, g_3, \ldots, g_k\}$ are defined according to Equation (1)) as follows:

$$\varphi(s) = \langle g_1(s), g_2(s), g_3(s), \ldots, g_k(s) \rangle \qquad \text{Equation (2)}$$

This state vector $\varphi(s)$ maps a concept s in the directed set to a point in k-space ($\mathbb{R}^k$). One can measure distances between the points (the state vectors) in k-space. These distances provide measures of the closeness of concepts within the directed set. The means by which distance can be measured include distance functions, such as those shown Equations (3a) (Euclidean distance), (3b) ("city block" distance), or (3c) (an example of another metric). In Equations (3a), (3b), and (3c), $\rho_1 = (n_1, p_1)$ and $\rho_2 = (n_2, p_2)$.

$$|\rho_2 - \rho_1| = (|n_2 - n_1|^2 + |p_2 - p_1|^2)^{1/2} \qquad \text{Equation (1a)}$$

$$|\rho_2 - \rho_1| = |n_2 - n_1| + |p_2 - p_1| \qquad \text{Equation (1b)}$$

$$(\Sigma(\rho_{2,i} - \rho_{1,i})^n)^{1/n} \qquad \text{Equation (1c)}$$

Further, trigonometry dictates that the distance between two vectors is related to the angle subtended between the two vectors, so means that measure the angle between the state vectors also approximates the distance between the state vectors. Finally, since only the direction (and not the magnitude) of the state vectors is important, the state vectors can be normalized to the unit sphere. If the state vectors are normalized, then the angle between two state vectors is no longer an approximation of the distance between the two state vectors, but rather is an exact measure.

The functions g_k are analogous to step functions, and in the limit (of refinements of the topology) the functions are continuous. Continuous functions preserve local topology; i.e., "close things" in S map to "close things" in $\mathbb{R}^k$, and "far things" in S tend to map to "far things" in $\mathbb{R}^k$.

Example Results

The following example results show state vectors $\varphi(s)$ using chain 1605 as function g_1, chain 1610 as function g_2, and so on through chain 1640 as function g_8.

φ("boy") $\Rightarrow$ ⟨3/4, 5/7, 4/5, 3/4, 7/9, 5/6, 1, 6/7⟩

φ("dust") $\Rightarrow$ ⟨3/8, 3/7, 3/10, 1, 1/9, 0, 0, 0⟩

φ("iguana") $\Rightarrow$ ⟨1/2, 1, 1/2, 3/4, 5/9, 0, 0, 0⟩

φ("woman") $\Rightarrow$ ⟨7/8, 5/7, 9/10, 3/4, 8/9, 2/3, 5/7, 5/7⟩

φ("man") $\Rightarrow$ ⟨1, 5/7, 1, 3/4, 1, 1, 5/7, 5/7⟩

Using these state vectors, the distances between concepts and the angles subtended between the state vectors are as follows:

Pairs of Concepts	Distance (Euclidean)	Angle Subtended
"boy" and "dust"	~1.85	~52°
"boy" and "iguana"	~1.65	~46°
"boy" and "woman"	~0.41	~10°
"dust" and "iguana"	~0.80	~30°
"dust" and "woman"	~1.68	~48°
"iguana" and "woman"	~1.40	~39°
"man" and "woman"	~0.39	~07°

From these results, the following comparisons can be seen:

- "boy" is closer to "iguana" than to "dust."
- "boy" is closer to "iguana" than "woman" is to "dust."
- "boy" is much closer to "woman" than to "iguana" or "dust."
- "dust" is further from "iguana" than "boy" to "woman" or "man" to "woman."
- "woman" is closer to "iguana" than to "dust."
- "woman" is closer to "iguana" than "boy" is to "dust."
- "man" is closer to "woman" than "boy" is to "woman."

All other tests done to date yield similar results. The technique works consistently well.

FIG. 1 shows a two-dimensional topological vector space in which state vectors are used to construct a semantic abstract for a document. (FIG. 1 and FIGs. 2 and 3 to follow, although accurate representations of a topological vector space, are greatly simplified for example purposes, since most topological vector spaces will have significantly higher dimensions.) In FIG. 1, the "x" symbols locate the heads of state vectors for terms in the document. (For clarity, the line segments from the origin of the topological vector space to the heads of the state vectors are not shown in FIG. 1.) Semantic abstract 105 includes a set of vectors for the document. As can be seen, most of the state vectors for this document fall within a fairly narrow area of semantic abstract 105. Only a few outliers fall outside the main part of semantic abstract 105.